Tutorial

Analysis of SARS-CoV-2 using MinION sequences

March 20, 2020

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
digitalinsights.qiagen.com · ts-bioinformatics@qiagen.com

Analysis of SARS-CoV-2 using MinION sequences

Sequencing using MinION is a fast and cost-efficient way to track the evolution of SARS-CoV-2 as it allows sequencing to be done in a few hours with little investment in laboratory equipment. However, samples are often of low quality and contaminated with RNA from other sources. Here we demonstrate how to overcome the challenge of metatranscriptomics and get the most from Oxford Nanopore MinION reads using the *CLC Genomics Workbench*.

The tutorial covers the following:

- Import of data required for the analysis.
- Trimming MinION reads.
- Mapping MinION reads to a reference.
- Calling variants in the sample relative to a reference and visualizing variant calls and mappings.
- Extracting a consensus sequence from a read mapping.
- Using BLAST to identify a strain.
- Running the pipeline in a workflow.

Prerequisites

For this tutorial, you must be working with *CLC Genomics Workbench* 20.0 or higher with the Long Read Support (beta) plugin installed.

How to install plugins using the Plugin Manager is described here: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Install.html

Plugins can also be downloaded from our website (https://digitalinsights.qiagen.com/products-overview/plugins/) and then installed using the Plugin Manager.

If you are not already familiar with the tools in Long Read Support (beta) plugin, we suggest first completing the tutorial "De Novo Assembly Using Long Reads and Short Read Polishing".

Background information

The data you will be working with is from metatranscriptome sequencing using a MinION data set from the Wuhan seafood market pneumonia outbreak [Chan et al., 2020]. Sampling the virus responsible for the disease involves obtaining sputum, throat or nasopharyngeal swabs or bronchoalveolar lavage fluid (BALF) samples from a patient. Because of this, samples will contain RNA from non-viral sources. This data set was prepared using the Sequence-Independent, Single-Primer Amplification (SISPA) protocol for additional viral sequence enrichment.

Some basic tips

- In this tutorial, we refer to the tools in the Workbench menus, but you can instead click on the **Launch** button () button in the top toolbar to search for and launch tools of interest.

- We will run several of the tools in "Batch mode". Using the "Batch" option allows us to launch the tool wizard once, but if several input elements are provided, the tool will run once per input element, analyzing each input element in turn, separately.

Import the data

1. Download the example data from our web site http://resources.qiagenbioinformatics.com/testdata/SARS-CoV-2_MinION_example_data.zip and unzip it.
2. Open the *CLC Genomics Workbench*.
3. Create a new folder for the project with a relevant name, for example, "SARS-CoV-2 MinION Tutorial".
4. Import the reference sequence and adapter list provided with the example data by going to:

   File | **Import** | **Standard Import**

   Ensure the option **Automatic import** is selected and select "MT135044.gbk" and "SISPA adapter trim list.clc". Save the items in "SARS-CoV-2 MinION Tutorial".

   After import, the SISPA adapter trim list should look like that shown in Figure 1.

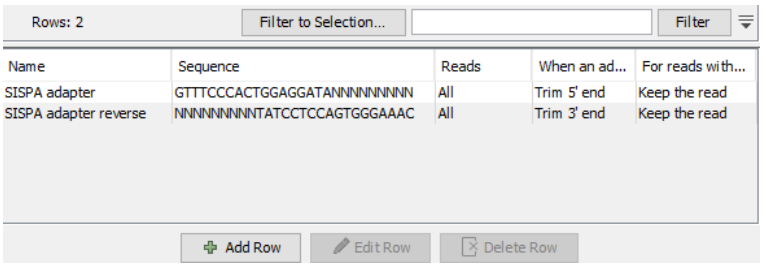

Figure 1: *Adapter trim list*

5. Download the full dataset the MinION reads and sample metadata using SRA.

   - Search for reads in SRA using **Download** | **Search for Reads in SRA (**  **)**.
   - Choose the term "Bioproject" in the drop down list in the top left side, and enter "PRJNA601630" into the search field.
   - In the results table, select the rows with Run accessions SRR10948474 and SRR10948550.
   - Click on the **Download Reads and Metadata** button, leave the options set as the defaults, and choose to save the results to a new folder, for example called "Raw Reads".

     Note: If, for some reason, you are unable to access SRA we have included the raw reads. To use these, go to **File** | **Import** | **Oxford Nanopore**.
     Click "**Add files** and select "SRR10948550.fastq" and "SRR10948474.fastq" then click **Next** and save to a new folder called "Raw Reads".

Information on these read sets

Run	# of Spots	# of Bases	Size	Sampling method
SRR10948474	505,484	284.6M	250.1Mb	Sputum
SRR10948550	425,717	146.3M	126.6Mb	Nasopharyngeal swab

Trim adapters from the reads

The samples were prepared using a SISPA protocol, so the first thing we will do is trim the reads using the **Trim Reads** tool and imported Trim Adapter Lists.

1. Run the **Trim Reads** tool to trim adapters from both sets of reads.
   - Go to:
     Toolbox | **Prepare Sequencing Data** () | **Trim Reads** ()
   - Select the sequence lists containing the reads and move these to the "Selected elements" field on the right. Click on **Next**.
   - Deselect both the options in the Quality trimming step and then click on **Next**.
   - In the Adapter trimming step of the wizard:
     - Click on the file selector icon to the right of the "Trim adapter list" field and then select the SISPA adapter trim list and click on **OK**.
     - Click on **Next**.
   - Leave options in the Trim homopolymers step and Sequence filtering steps unchecked and click on **Next** in each case.
   - In the Results handling steps, select the **Create report** option and choose to save the output to a new folder, for example named "Trimmed reads".

Map long reads to reference

In this section, you will map the reads to a closely related strain to create a new consensus sequence. We will use isolate MT135044 as the reference. This was imported in an earlier step. We stress that this is an example, and any high-quality assembly of a closely related strain will work.
Note: Since we are dealing with metatranscriptomic data, we do not expect all reads to map and coverage may be quite low as a result.

1. Run the **Map Long Reads to Reference (beta)** by going to:
   Toolbox | **Long Read Support (beta)** () | **Map Long Reads to Reference (beta)** ()
2. Select the trimmed reads created in the previous step. Check **Batch** to run once per sample and click **Next**.
3. Leave the Batch overview settings as is and click **Next**.
4. In references, use () to select the MT135044 reference sequence you imported and click **OK** and **Next**.
5. Leave the mapping options on default and click **Next**.
6. Check the **Create report** option. Click **Next** and **Save** a new folder "Read mappings to reference". You can monitor the progress of the read mapping job under the Processes tab in the bottom left of the Workbench. The tool will take approximately 1-2 minutes per sample depending on your hardware.

7. After the tool has finished running, inspect the mapping reports. The table below shows an example of the read mapping statistics:

Read mapping report

Run	Count	% reads	Average Length	# of bases	% bases
SRR10948474	191,135	37.91%	663.06	126,734,732	62.34%
SRR10948550	996	0.23%	702.25	699,445	0.80%

It can be seen that a large percentage of reads in SRR10948550 are not from SARS-CoV-2.

Call variants relative to the reference sequence

1. Now call variants on these read mappings using the Fixed Ploidy Variant Detection tool.

   Resequencing Analysis ( ) | **Variant Detectors** ( ) | **Fixed Ploidy Variant Detection** ( )

2. Select the read mappings and choose to run the job in batch mode.

3. Change the Fixed ploidy variant options to **Ploidy** = 1, **Required variant probability (%)** = 80. Click **Next**.

4. Set **Minimum coverage** = 10, **Minimum count** = 7, **Minimum frequency(%)** = 75.0. Click **Next**.

5. Disable all options in Noise filters. Click **Next** and save the variant calls in a new folder "SARS-CoV-2 variant calls".

6. Open the output variant tracks to observe the result. Both runs have variant calls relative to the MT135044 reference, three in common with each other. Below is an example of the SRR10948474 variant calls:

SRR10948474 variant calls

Chromosome	Region	Type	Reference	Allele
MT135044	4402	SNV	C	T
MT135044	5062	SNV	T	G
MT135044	9561	SNV	C	T
MT135044	15607	SNV	T	C
MT135044	29095	SNV	C	T

Similarly, an example of the SRR10948550 variant calls is displayed below:

SRR10948550 variant calls

Chromosome	Region	Type	Reference	Allele
MT135044	4402	SNV	C	T
MT135044	5062	SNV	T	G
MT135044	29095	SNV	C	T

The sampling dates are 28th January 2020 for MT135044 and 11th January for SRR10948474 and SRR10948550. This shows the ability to trace mutations over time during an active outbreak.

7. Now the variants and read mappings can be inspected together by selecting read mappings ( ) and variant tracks ( ) for all runs and execute **File** | **New** | **Track List** ( ). *Tip*: As MinION reads have many errors, a more stringent display of the read mapping gives a

better experience. To achieve this, change the Track List Settings - Track layout - Reads track - Hide insertions below (%) to e.g. 80.

8. Optional: You can include amino acid changes by first creating a genome and CDS track.
   - To start, open **Track Tools ( )** | **Track Conversion** | **Convert to Tracks ( )**
   - Select MT135044 as input and click **Next**.
   - Check **Create sequence track** and **Create annotation tracks**. In **Annotation types**, press ( ), select CDS and click **Done** to close the dialog. Then click **Next**.
   - Click **Next** and save to a location titled "Tracks" and click **Finish**.
   - Next, annotate variants and create a track for visual inspection by running the **Resequencing Analysis ( )** | **Functional Consequences** | **Amino Acid Changes ( )** tool.
   - Select the variant tracks and remember to check to run in batch mode. Click **Next**.
   - In **Set parameters** select the CDS and Genome tracks created in previous steps as shown on Figure 2. Leave the other settings on default. Click **Next**, then save to your variant location.

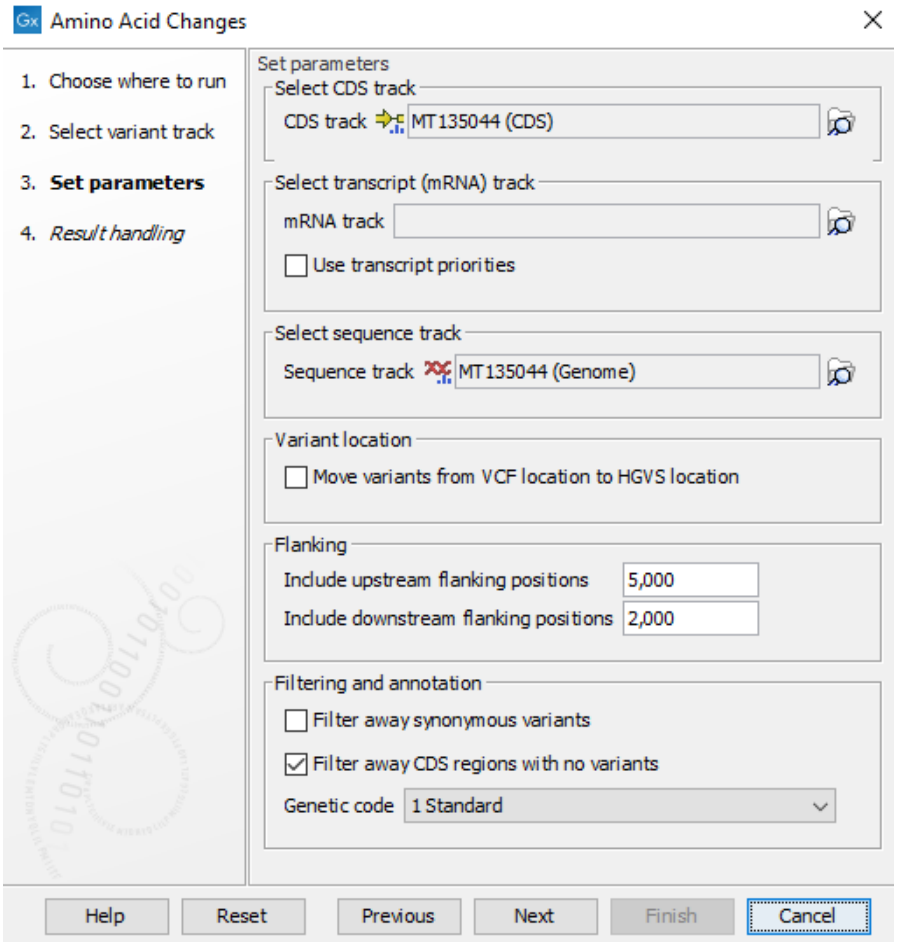

Figure 2: *Amino Acid Changes options*

   - You can now drag and drop the Amino Acids tracks into the Track Viewer.

9. Your track list should look like Figure 3.

Figure 3: *Read mapping of MinION samples and variant call visualization*

Extract consensus sequences from the read mappings

We have shown how to use an existing reference to call variants. This section focuses on how to use the read mapping to construct consensus sequences for these samples. The consensus sequences will be used in the next steps to search for similar sequences in the NCBI Betacoronavirus BLAST database.

1. Run the **Resequencing Analysis ( )** | **Extract Consensus Sequence ( )** tool. Select the read mappings created in previous step and check **Batch** to run once per sample. Click **Next**.

2. Leave **Low coverage definition** and **Low coverage handling** set to their default values and click **Next**.

3. Check **Insert ambiguity codes** and set **Noise threshold** = 0.7 and **Minimum nucleotide count** = 1.
   We raise the noise threshold to prevent homopolymeric regions from creating noise in our assembly. Your options screen should look like Figure 4. Click **Next**.

4. Select to save the consensus sequences in a new folder titled "Consensus sequences".

5. Lastly, run **Utility Tools** | **Batch Rename** tool. We do this to keep it easier to keep track of our sequences in the later steps.

   - Select the consensus sequences and click **Next**.
   - Check **Rename sequences in sequence lists** as shown in Figure 5

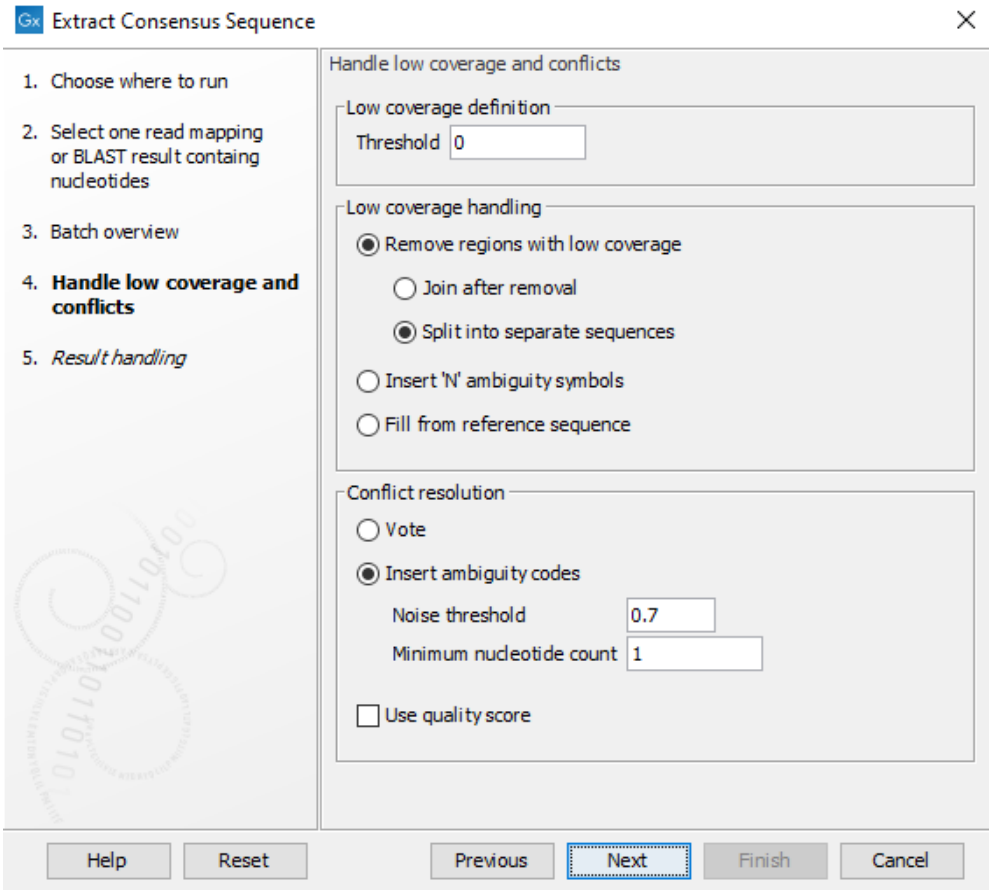

Figure 4: *Extract Consensus Sequence options*

- Select the **Replace full name** option, and enter "#BR-PE#".
  This will change the sequence names to match the parent filenames. This is shown in Figure 6.

You should now have two complete, high quality sequences.

Using BLAST to compare to other SARS-CoV-2 assemblies

The sequences can be used as queries in a BLAST search to see which assemblies they match. We will use a BLAST database from the NCBI as this is publicly available. However, we recommend using more complete databases when analysing your own data, for example GISAID if possible.
To download a BLAST database from the NCBI, your Workbench must be able to connect to the internet.

1. Download a Betacoronavirus BLAST database from the NCBI by going to:

   Toolbox | **BLAST** ( )| **Download BLAST Databases** ( )

2. Select the Betacoronavirus database from the list of databases available and then click on **Finish**. Now, run BLAST once for each consensus sequence we generated in the previous step:

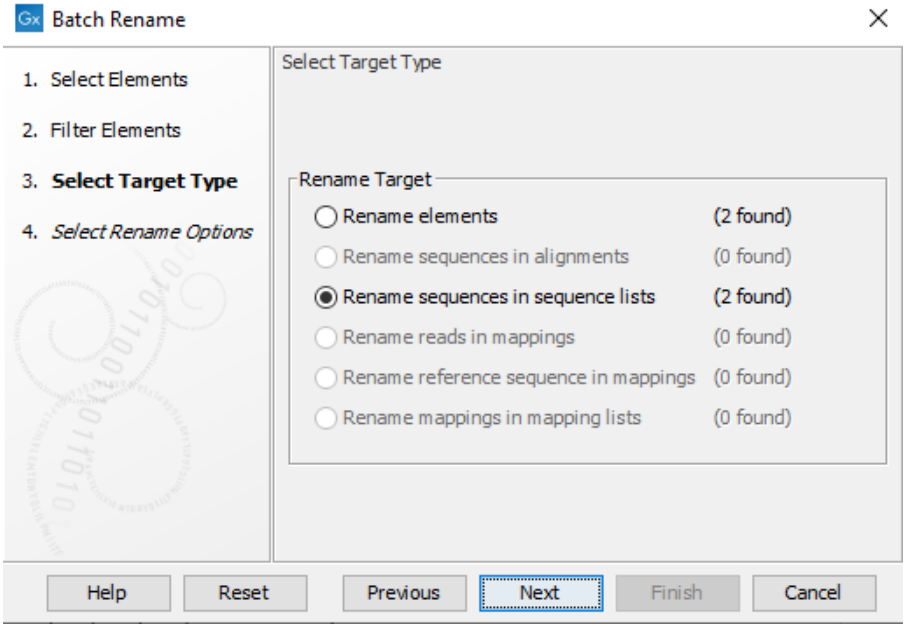

Figure 5: *Select to rename sequences*

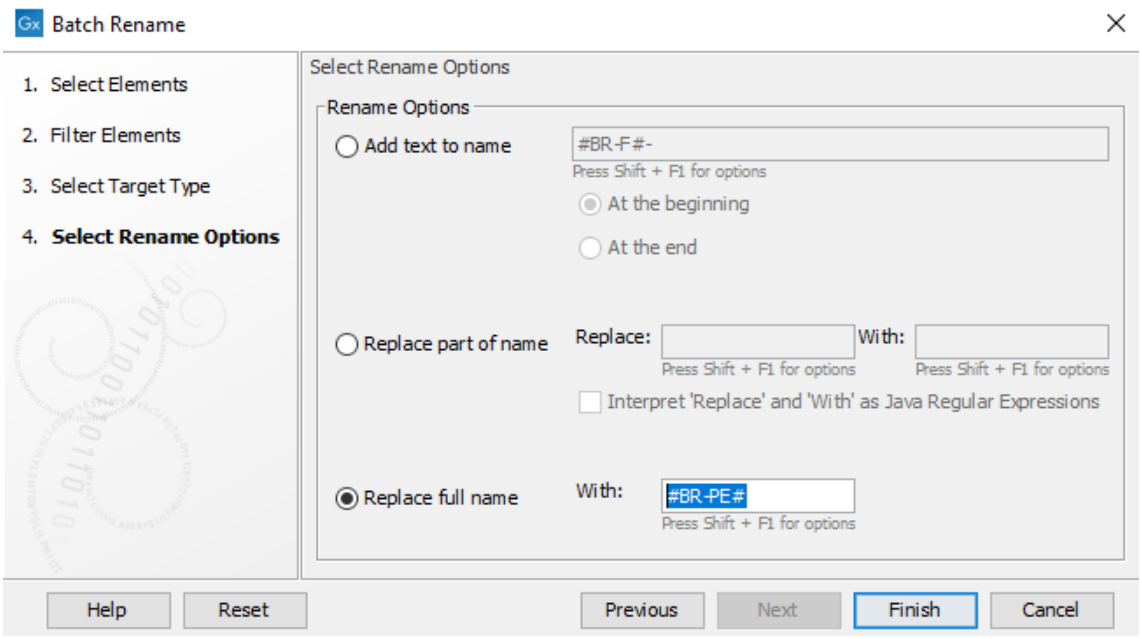

Figure 6: *Replace full name with parent filename*

3. Launch BLAST by going to:

   Toolbox | **BLAST** ( )| **BLAST** ( )

4. Put the sequence lists generated into the Selected elements area and click on **Next**.

5. Choose "blastn" as the type of search to run and select the option "Blast database" as the Target type. From the drop down options then available, select the "Betacoronavirus (DNA) - Betacoronavirus" database and then click on **Next**.

6. We are looking for sequences similar to our query sequence and are only interested in the

top hits, so change **Expect**= 0.01 and leave the default search settings and click on **Next**.

7. Choose to "Open" the results rather than saving them, and click on **Finish**.
8. From the results view, take a note of the best matching sequences in the database. The "Show BLAST HSP Table" includes values such as %Identity, %Positive, %Gaps. The results can be sorted by click on the column names. Among the top hits for the consensus sequence for SRR10948550, we expect to see MN938384. For the consensus sequence from SRR10948474, we expect to see MN975262. These are the sequences submitted by [Chan et al., 2020] that were assembled based on the same read sets we are using.
9. You can download the reference of your BLAST result by selecting the BLAST hit of SRR10948474 in the results accession column.
    - Click on the "Open BLAST Output" button.
    - In the opened BLAST output, change the view to the second tab "Show BLAST hit table". Select MN9383384 and click **Download and Save** (Figure 7).

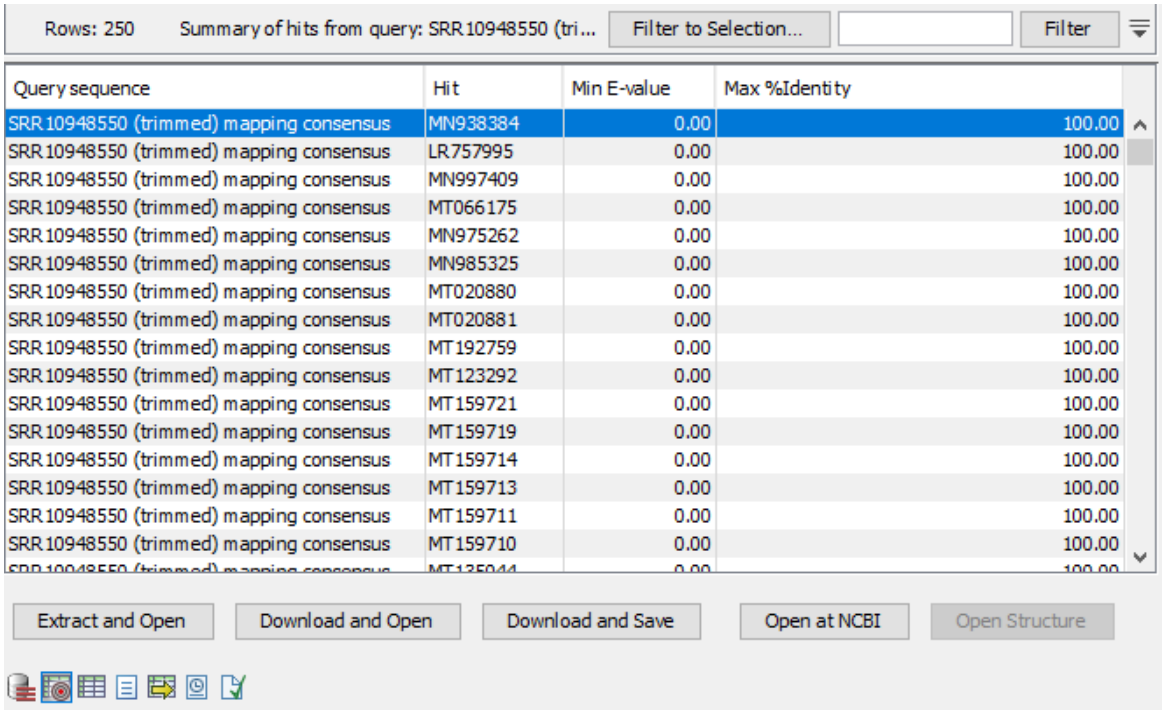

Figure 7: *BLAST output hit table*

    - Save this new sequence in a new folder, for example titled called "Best matched references".
10. Repeat the above steps for SRR10948550. You have now identified the known strains most closely related to your samples.
11. Note: You can also create your own BLAST database. To do this, import or download the sequences you wish to include in the database. Then launch **Create BLAST Database** tool.

Run the analysis using a workflow

1. To automate the process for reproducible execution of the trimming, mapping, variant calling and consensus sequence extraction, we have provided a workflow in the zip file above. The steps to rename, download a BLAST database and BLAST have to be executed manually.

2. Workflows can be installed using Workflow manager. An installed workflow can then be found in the **Toolbox** under **Installed workflows**.

3. You can inspect the workflow by double-clicking on it, and you can run it on any number of samples in batch mode. On the MinION reads provided in the introduction, this workflow takes less than 10 minutes on a 2013 MacBook pro laptop computer (2,6 GHz Intel Core i5).

Figure 8 shows the automated pipeline of the bioinformatics steps in the workflow editor.

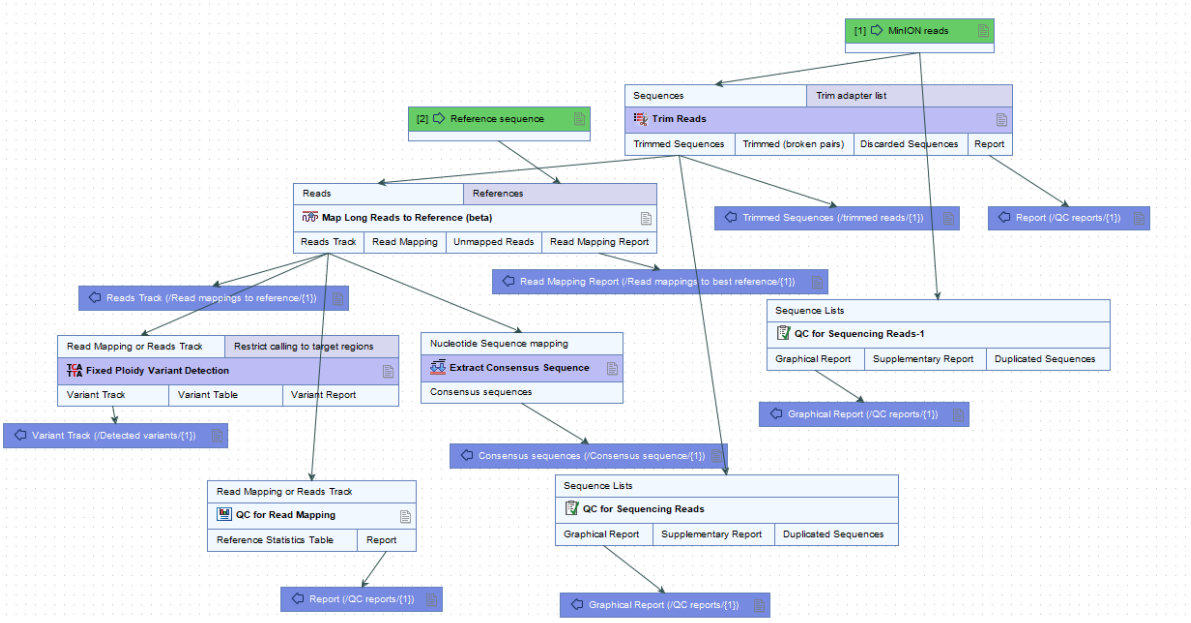

Figure 8: *A workflow reproducing the bioinformatics pipeline*

Bibliography

[Chan et al., 2020] Chan, J., Yuan, S., Kok, K., To, K., Chu, H., Yang, J., Xing, F., Liu, J., Yip, C., Poon, R., Tsoi, H., Lo, S., Chan, K., Poon, V., Chan, W., Ip, J., Cai, J., Cheng, V., Chen, H., Hui, C., and Yuen, K. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet*, 395(15):514–523.